2-D Processing of Speech
with Application to Pitch and Formant Estimation*

Thomas F. Quatieri and Tianyu Tom Wang
MIT Lincoln Laboratory

Harvard Workshop on
Next-Generation Statistical Models for
Speech and Audio Signal Processing

November 9-10 2007

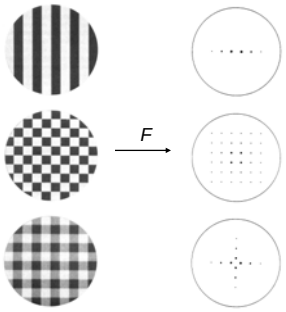
*This work was supported by the Department of Defense under Air Force contract FA8721 05 C 0002. The opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government..

Motivation
From Image Processing

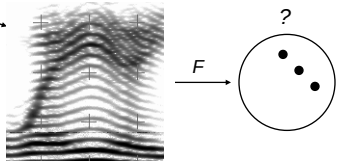
- **Certain 2-D geometric patterns transform to dots in a 2-D spatial frequency plane***
- **Time-frequency distributions contain "geometric patterns" due to harmonic content**
  - **Possible use**
    - **Pitch estimation**
    - **Noise reduction**
    - **Multi-speaker separation**

**From R.L. DeValois and K.K. DeValois, *Spatial Vision*, Oxford University Press, 1988.*

2-D Grating Patterns



Narrowband Spectrogram



Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
2-D Processing of Speech with Application to Pitch and Formant Estimation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Lincoln Laboratory,244 Wood Street,Lexington,MA,02420-9108**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

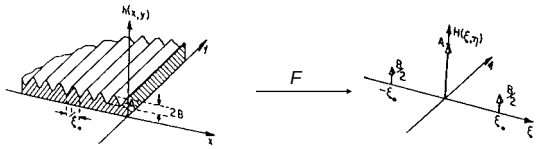
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**14**	

OUTLINE

- **2-D Spectrogram Model and Mapping**
- **Application to Pitch Estimation**
- **Application to Formant Estimation**
- **Extension to Alternate Time-Frequency Distributions**
- **Conclusions and Directions**

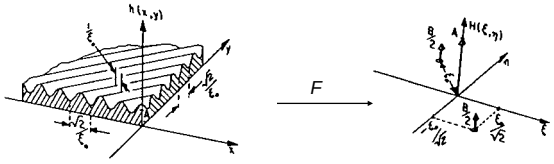
2-D Spectrogram Model
Inspiration from Image Processing

- **2-D sine on a pedestal*: Zero degree rotation**



Distance of 2-D impulses from origin varies inversely with sine frequency

- **2-D sine on a pedestal*: 45 degree rotation**



Angle of 2-D impulses w/r axes proportional to extent of sine rotation

From J.D. Gaskill, *Linear Systems, Fourier Transforms, and Optics*, John Wiley and Sons, New York, NY, 1978.

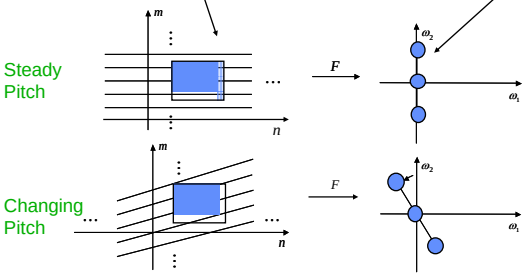
2-D Spectrogram Model

Short-Space 2-D Sine

- **Harmonic line structure of the narrowband spectrogram is modeled over a <u>small region</u> by a 2-D sine function sitting on a flat pedestal of unity**
- **2-D window is applied to extract a short-time segment and 2-D Fourier transform is then computed**

$$x[n,m] = w[n,m](1+\cos(\omega_g m))$$

$$X(\omega_1,\omega_2) = 2W(\omega_1,\omega_2) + W(\omega_1,\omega_2-\omega_g) + W(\omega_1,\omega_2+\omega_g)$$



2-D Impulses are smeared by 2-D windowing

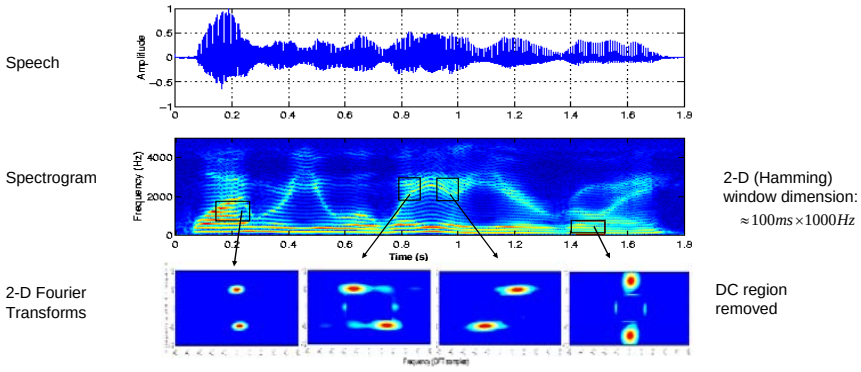
Distance from origin of smeared impulses varies inversely with pitch

Angle of impulses from vertical axis varies with pitch dynamics

2-D Processing

Example

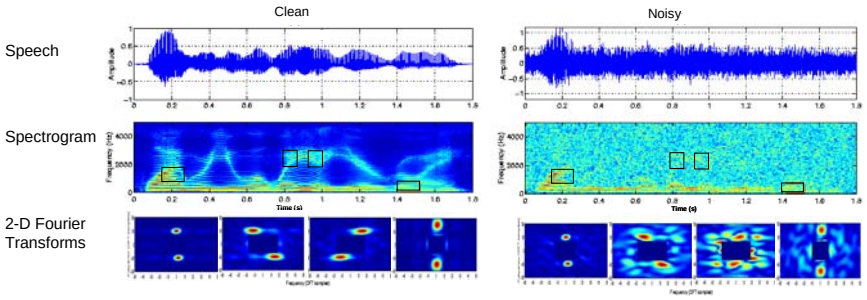
- **2-D analysis of narrowband spectrogram of all-voiced female speech**



Henceforth, refer to 2-D mapping as the "<u>G</u>rating <u>C</u>ompresssion <u>T</u>ransform" (GCT) to highlight mapping "gratings" to concentrated "dots"

2-D Processing
Example with Noise

- **2-D analysis of all-voiced female speech in noise**
  - **GCT without and with additive white Gaussian noise at average SNR of ~3 dB**



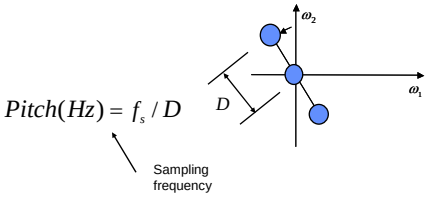
- **Energy concentration of GCT is typically preserved at roughly the same location as for the clean case**
  - **However, when noise dominates so that little harmonic structure remains within the 2-D window, energy concentration deteriorates, as in the vicinity of 0.95 s and 2000 Hz**

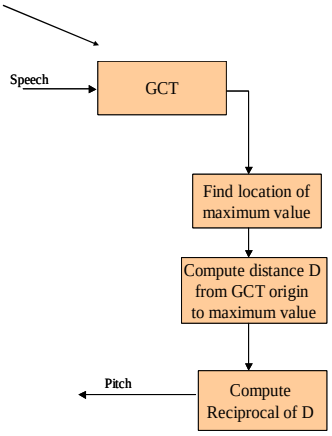
OUTLINE

- **2-D Spectrogram Model and Mapping**
- **Application to Pitch Estimation**
- **Application to Formant Estimation**
- **Extension to Alternate Time-Frequency Distributions**
- **Conclusions and Directions**

Pitch Estimation
GCT-Based Approach

- GCT of speech examples motivate a simple pitch estimator
  - Pitch estimate is reciprocal to the distance from the origin to the maximum value in the GCT

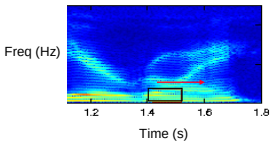




- Pitch rate of change is proportional to angle of GCT peak from vertical axis

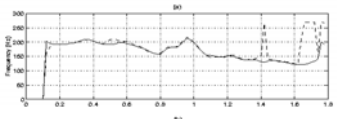
GCT-Based Pitch Estimation
Example

- GCT-based estimator over time
  - 2-D analysis window slid along the speech spectrogram at a 10-ms frame interval at low-frequency location

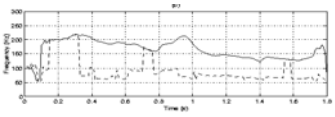


  - Relatively robust in noise, out-performing a sinewave-based pitch estimator

GCT-Based Estimator



Sinewave-Based Estimator



Solid: Clean Speech
Dashed: 3 dB SNR

GCT-Based Pitch Estimation
Performance

- **GCT-based estimator**
  - **2-D analysis window slid along the speech spectrogram at a 10-ms frame interval at a low-frequency location given by the 2-D window in previous slide**
  - **Average magnitude difference measured between pitch-contour estimates with and without white Gaussian noise for both the GCT- and sinewave-based estimators**

Performance Measurements

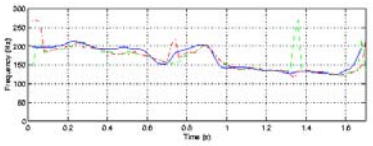
	FEMALES		MALES	
	9dB	3dB	9dB	3dB
GCT	0.5	6.7	0.9	6.7
SINE	5.8	40.5	2.6	12.8

Average magnitude error (in dB) in GCT- and sine-wave-based pitch contour estimates for clean and noisy all-voiced passages. The two passages "Why were you away a year Roy?" and "Nanny may know my meaning." from two male and two female speakers were used under noise conditions 9 dB and 3 dB average signal-to-noise ratio.

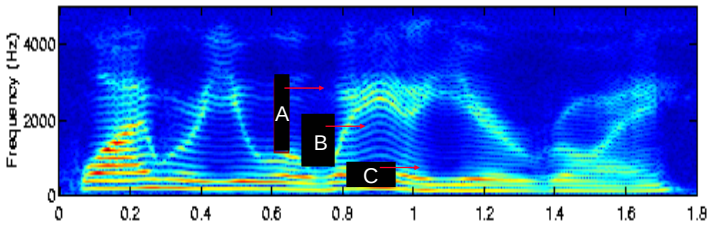
Pitch Estimation
Multi-resolution properties

- **Pitch in the 2-D plane**
  - **Pitch can be obtained anywhere in the 2-D plane**
  - **"Wavelet-like tiling" of 2-D window found to give the most consistent estimate**
    - **Reflects increase pitch FM with increasing frequency**

Three pitch contours

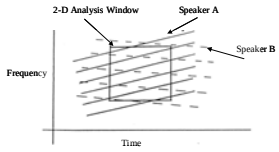


Window A: Dashed Green
Window B: Dashed-Dot Red
Window C: Solid Blue

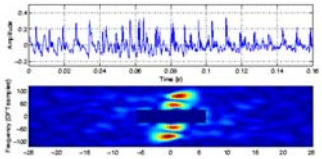


Two-Speaker Pitch Estimation

- Sum of two speakers has spectrogram with two harmonic sets

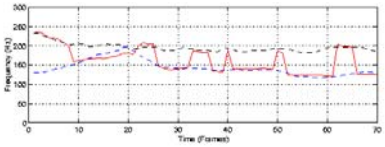


- GCT gives two pairs of dots, one pair for each speaker
  - All-voiced example (male + female)



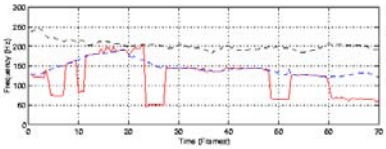
- Blind use of one-speaker pitch estimator on two-speaker signal

GCT-Based Estimator



Maximun of GCT latches on to speaker with (locally) largest energy

Sine-Wave-Based Estimator



OUTLINE

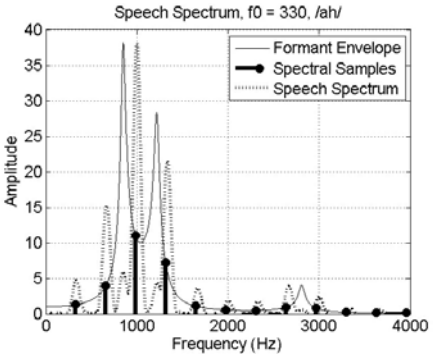
- 2-D Spectrogram Model and Mapping
- Application to Pitch Estimation
- Application to Formant Estimation
- Extension to Alternate Time-Frequency Distributions
- Conclusions and Directions

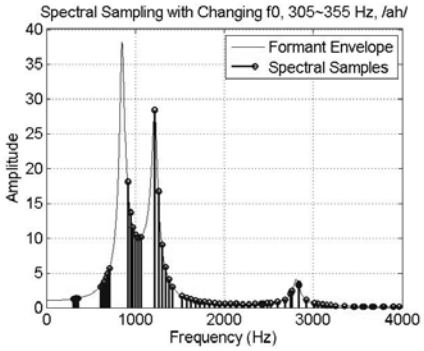
Formant Estimation

The High-Pitched Problem

Synthesized vowel /ah/ with 330-Hz pitch. Speech spectrum generated from short-time Fourier analysis with a 20-ms Hamming window.

Collection of harmonic samples from pitch sweep ranging from 305~355 Hz. Contrast to *f0* = 330 Hz shown in Figure 1.

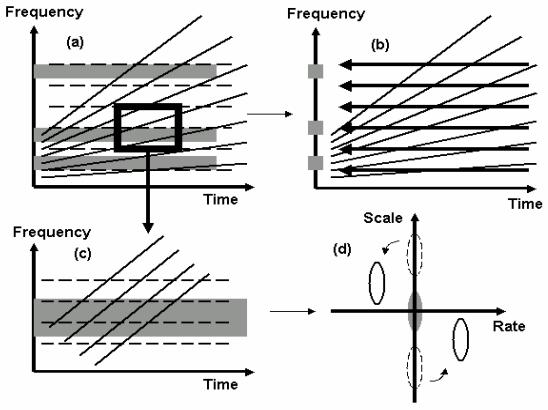




2-D Framework

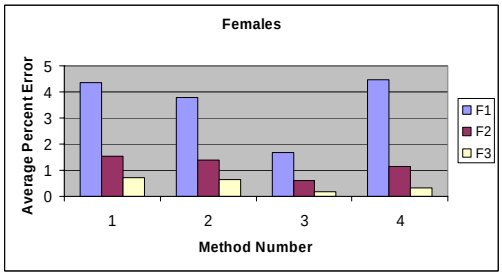
Exploiting Changing Pitch in Formant Estimation

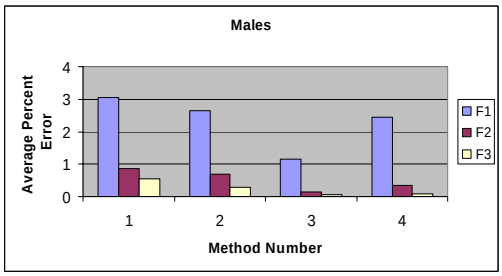
a) Schematic of changing and fixed *f0* across a steady vowel in a STFT; (b) Averaging of harmonic samples to a single 1-D frequency axis; (c) Localized spectrotemporal region from (a); (d) Mapping of source-filter speech components in the GCT from (c).



Spectral Estimation: Results

Average Percent Formant Error

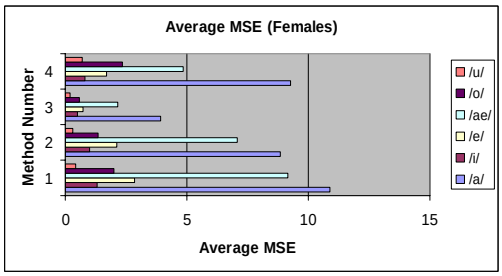


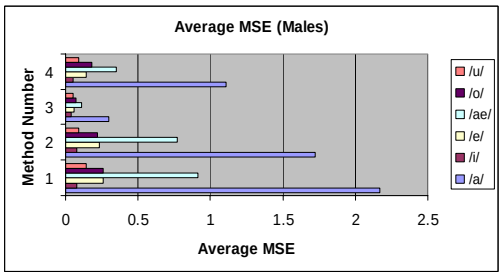


- **Methods**
  1. **Single STFT slice**
  2. **Cepstral liftering**
  3. **Proposed method**
  4. **Spectral slice averaging**
- **Relative gains (method 1) for [F1, F2, F3] via proposed:**
  - **Females: [61%, 61%, 73%]**
  - **Males: [62%, 82%, 87%]**
- **Gains for F3 greatest (wider harmonic sampling)**
- **Individual formant scores *across* vowels also consistent**

Spectral Estimation: Results

Average MSE



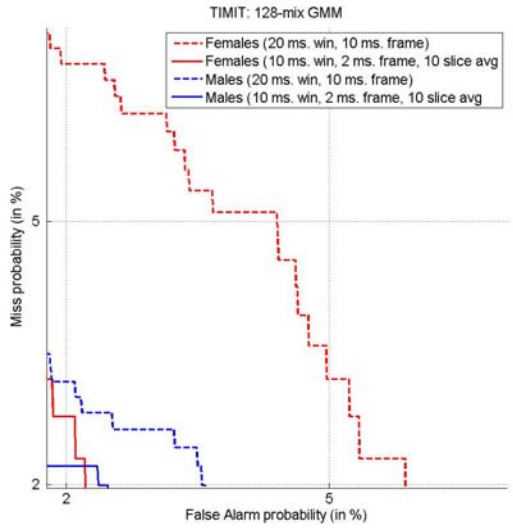


- **Methods**
  1. **Single STFT slice**
  2. **Cepstral liftering**
  3. **Proposed method**
  4. **Spectral slice averaging**
- **Results are consistent with formant frequency estimation**
  - **Method 3 outperforms others for all vowels**
- **Data not shown:**
  - **Consistent results with children's formants**

Speaker Recognition: Methods

- Data set - male and female subsets of TIMIT corpus
- Baseline system
  - Mel-cepstrum feature extraction with 20 ms. window and 10 ms frame interval + delta features
  - Adaptive Gaussian Mixture Modeling
    - 128 mixture components
    - Universal background model
- System modifications in *feature extraction*
  - Short-time analysis using 10 ms frame window and 2 ms frame interval
  - Compute *average* of spectral slices spanning ~30 ms
  - Derived spectra are used for computing standard mel-cepstrum + deltas

Speaker Recognition: Results



- Equal error rate (EER)

	Baseline (EER)	Proposed (EER)
Males	1.86%< 2.45% < 3.39%	1.53% < 2.15% < 2.80%
Females	3.12%< 4.41% < 5.64%	1.55% < 2.15% < 3.30%

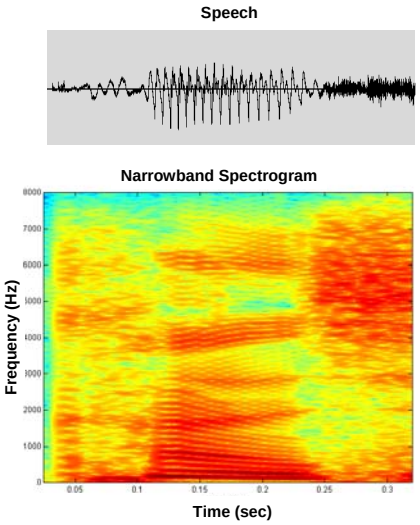
- Absolute EER reduction in females 2.26% (not yet significant)

OUTLINE

- 2-D Spectrogram Model and Mapping
- Application to Pitch Estimation
- Application to Formant Estimation
- Extension to Alternate Time-Frequency Distributions
- Conclusions and Directions

Limitation of the Spectrogram
Observations

- Two curious effects are seen:
  - Frequency tracks moving in the wrong direct, e.g., up rather than down and
  - Crossing tracks, i.e., tracks moving up and down simultaneously.
- The problem is that the basis functions of the Fourier transform, stationary sinusoids, cannot resolve the speech harmonics which have rapid frequency modulation and are closely spaced in frequency.
  - This lack of resolution leads to the complex line phenomena seen in Figure 2.

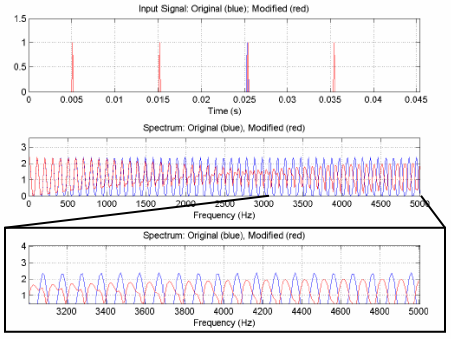


Spectral Sensitivity*

Example

- Harmonic speech spectra can be quite sensitive to aberrations in periodicity of the glottal source
- Even small perturbations can lead to short-time spectral changes that mislead the viewer in terms of signal composition

Example: One-sample shift (0.1 ms)



We see that the shift in time domain seems to move the harmonics in the higher frequencies

*We have developed formal spectral for these sorts of effects: To be published in January 2008 IEEE TSLP, "Spectral representations of nonmodal phonation," Malyska and Quatieri.

An Alternate Transform

The Fan-Chirp Transform

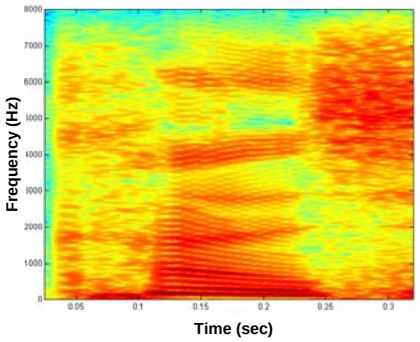
- The Fan-Chirp Transform (FChT)
  - "Adaptive Chirp-Based Time–Frequency Analysis of Speech Signals"
    Marian Kepesia and Luis Weruaga, *Speech Communication*, vol. 48, no. 5, pp. 474-492, May 2006.
  - "The Fan-Chirp Transform for Non-Stationary Harmonic Signals"
    Luis Weruaga and Marian Kepesia, (submitted to Elsevier)
- FChT is a generalization of the Fourier transform
  - Fourier transform basis functions are stationary sine waves
  - FChT basis functions are sine waves with linear frequency modulation
    - the set of basis functions has a fan geometry
    - 1st order match to harmonic frequency modulation in speech

Spectrogram Comparison

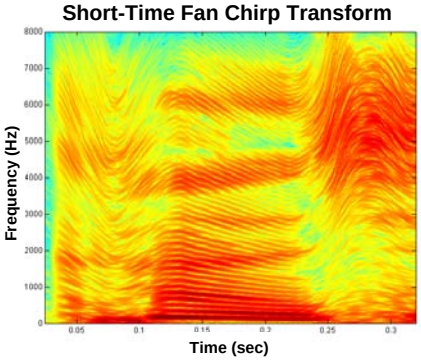
STFT versus Short-Time Fan-Chirp

- **Observations**
  - **FChT resolves high frequency harmonics even when frequency modulation is large**
  - **Frequency tracks appear as predicted for FChT**

Short-Time Fourier Transform



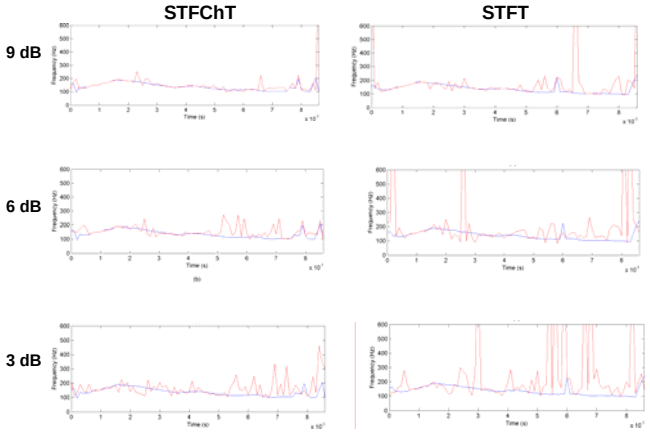
Short-Time Fan Chirp Transform



Fan-Chirp for Grating Compression Transform

Pitch Estimation

- **Based on preliminary results, for pitch estimation in noise, the short-time Fan-chirp tranform appears to outperform the STFT**



OUTLINE

- 2-D Spectrogram Model and Mapping
- Application to Pitch Estimation
- Application to Formant Estimation
- Extension to Alternate Time-Frequency Distributions
- Conclusions and Directions

Conclusions and Directions

- The grating compression transform (GCT) maps harmonically-related signal components to a concentrated entity in a spatial 2-D frequency plane
- The GCT forms the basis of a pitch estimator that uses the radial distance to the largest peak of the GCT
  - The resulting pitch estimator appears robust under noise conditions and amenable to extension to two-speaker pitch estimation
- The GCT forms the basis of a formant estimator that exploits separability of speech source and vocal tract information via changing pitch
- Although the spectrogram provides a useful starting point for the GCT, alternate transforms can provide improved performance
  - Fan-chirp transform is one possibility
- Possible GCT directions
  - Alternate time-frequency distributions
  - Pitch estimation
    - Extended evaluation to a larger corpus and use of voiced/unvoiced speech
    - Two-speaker pitch estimation
  - Formant estimaiton in noise
  - GCT as model of auditory cortical processing (Sthamma, Ezzat, and Poggio)